

# AI 보안과 딥페이크 기술 동향 강연 최종 요약문

## □ 핵심 내용

- AI 기술의 급속한 대중화 속에서 프롬프트 인젝션부터 딥페이크까지 다층적 보안 위협이 현실화되고 있으며, AI 시스템 자체의 보안과 AI를 악용한 허위정보 대응이라는 이중 과제의 체계적 대응 방안을 제시

## □ 배경 및 현황

- 초등학생들이 AI 이미지 생성 도구를 자연스럽게 활용하는 등 AI 기술이 전 연령층으로 확산되며 일상 기술로 자리잡음
- 2023년 103개에서 2024년 29만개로 급증한 텍스트 생성 모델(허깅페이스 기준), 가짜 댓글 생성 비용 1개당 약 1원으로 대규모 허위정보 생산 가능
- 언어모델의 근본적 한계(문맥 기반 단어 예측)와 학습 데이터 내 위험 정보(폭탄 제작법, 개인정보 등) 포함으로 구조적 취약점 내재

## □ 제안 방안

- 입력 단계 보안:** Meta의 Llama Guard를 활용한 프롬프트 필터링으로 폭력, 증오, 성적 콘텐츠 등 위험 요소 사전 차단
- 시스템 레벨 대응:** AI DLP(Data Loss Prevention) 도입으로 입출력 통제, AI 레드팀 운영을 통한 선제적 취약점 점검
- 탐지 기술 고도화:** 공간적 특징(손가락 개수, 눈동자 방향), 주파수 영역 분석, AI 기반 탐지 모델 개발로 딥페이크 식별
- 워터마킹 기술:** Google SynthID(이미지), Gemini(텍스트) 방식의 생성 단계 워터마크 삽입으로 AI 생성물 추적

## □ 주요 효과 및 장점

- 다중 방어 체계 구축:** 프롬프트 인젝션 공격 차단률 향상(딥시크 100% → GPT 80%대로 개선 가능)
- 실시간 위협 대응:** AI 에이전트를 활용한 24시간 위협 탐지 및 자동 대응 시스템 구현
- 투명성 확보:** 워터마킹을 통한 AI 생성 콘텐츠 식별로 진위 판별 가능(DALL-E, 삼성 갤럭시 지우개 적용 사례)
- 비용 효율성:** 네이버 하루 댓글량(20만개) 기준 가짜 댓글 생성 비용 20만원 대비 탐지 시스템 구축으로 방어 가능

## □ 핵심 인사이트 및 시사점

- 양면성 인식:** AI 기술이 시각장애인 지원 서비스부터 무기 시스템 공격까지 활용되는 이중성을 인지하고 균형있는 접근 필요
- 선제적 대응:** 우크라이나 전쟁 정보전, 엠버스방 성착취물, 국방부 피싱메일 등 실제 피해 사례를 통해 사후 대응보다 예방 중심 전략 수립 중요
- 국제 협력 필요성:** 딥시크 사례(중국 서버, 탈옥 성공률 100%)에서 보듯 글로벌 차원의 AI 보안 표준과 협력 체계 구축 시급
- 기술-규제 균형:** MITRE ATLAS, OWASP AI Top 10 등 표준 프레임워크 활용과 함께 혁신을 저해하지 않는 스마트 규제 필요

## □ 용어 정리

- **프롬프트 인젝션**: AI에 악의적 명령을 주입하여 제한을 우회하거나 민감정보를 유출시키는 공격 기법
- **할루시네이션(Hallucination)**: AI가 학습하지 않은 내용을 그럴듯하게 생성하는 현상(예: 세종대왕 맥도날드 사건)
- **탈옥(Jailbreaking)**: 장문의 프롬프트로 AI 모델을 가스라이팅하여 안전장치를 해제하는 기법
- **워터마킹**: AI 생성 콘텐츠에 비가시적 표식을 삽입하여 출처를 추적 가능하게 하는 기술
- **AI 레드팀**: 조직 내에서 AI 시스템의 취약점을 선제적으로 찾아 테스트하는 보안 전문 팀